

Faculty of Engineering and Information Technology  
University of Technology, Sydney

# **Contrast Mining in Large Class Imbalance Data**

A thesis submitted in partial fulfillment of  
the requirements for the degree of  
**Doctor of Philosophy**

by

Jinjiu Li

May 2013

## **CERTIFICATE OF AUTHORSHIP/ORIGINALITY**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

---

# Acknowledgments

First and foremost, I would like to express the deepest appreciation to my supervisor, Professor Longbing Cao, for providing me with all the necessary facilities and his great effort in leading me to the spirit of freedom in regard to research. Without his guidance and persistent help this thesis would not have been possible.

Thanks to my co-workers, Can Wang and Wei Wei, for being so supportive especially when I was struggling through hard times.

I am also grateful to my dear friends and colleagues in AAI, especially fellow students Mu li and Chunming Liu for their hard working in my project teams.

I also place on record, my sense of gratitude to the team members in Westpac for their expert and sincere help in the project.

I would like to express my eternal appreciation towards my family, my wife, my daughter Jessie and my son Vincent. Without their unconditional love, I could not have outlasted those tough times.

Jinjiu Li

December 2012 @ UTS

# Contents

Certificate . . . . .	i
Acknowledgment . . . . .	ii
List of Figures . . . . .	viii
List of Tables . . . . .	xi
Abstract . . . . .	xii
<b>chapter 1 Introduction . . . . .</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Research Issues & Significance . . . . .	3
1.2.1 Contrast Pattern Mining on Class Imbalance Data . . .	3
1.2.2 Predicative Rule Selection . . . . .	7
1.2.3 Personalized Domain Driven Feature Mining . . . . .	14
1.3 Main Research Objectives . . . . .	17
1.4 Research Contribution . . . . .	18
1.4.1 Personalised Domain Driven Feature Mining . . . . .	18
1.4.2 Contrast Pattern Mining of Class Imbalance data . . .	19
1.4.3 Globally Optimal Predicative Rule Selection . . . . .	20
1.4.4 Anomaly Detection System:i-Alertor & i-Educator . . .	21
1.5 Thesis Organization . . . . .	21
<b>chapter 2 Related Work . . . . .</b>	<b>23</b>
2.1 Introduction . . . . .	23
2.2 Frequent Pattern Based Classification . . . . .	24
2.2.1 Association Rule Mining . . . . .	24

2.2.2	Classification Rule Mining . . . . .	28
2.2.3	Associative Classification . . . . .	30
2.3	Discrimination Based Classification . . . . .	32
2.3.1	Sporadic Rules . . . . .	32
2.3.2	Contrast Set . . . . .	33
2.3.3	Subgroup Discovery . . . . .	34
2.3.4	Emerging Pattern & Jumping Emerging Pattern . . . . .	34
2.4	Feature Mining . . . . .	38
2.4.1	Feature Extraction . . . . .	38
2.4.2	Feature Selection . . . . .	40
2.5	Classification in Class Imbalance Data . . . . .	41
2.5.1	Data-Level Approaches . . . . .	42
2.5.2	Algorithm-Level Approaches . . . . .	42
2.6	Summary and Conclusion . . . . .	43
<b>chapter 3</b>	<b>Personalised Domain Driven Feature Mining . . . . .</b>	<b>45</b>
3.1	Introduction . . . . .	45
3.2	Problem Definition . . . . .	48
3.3	Operators for Feature Mining . . . . .	50
3.4	Learning Application of Operators . . . . .	52
3.4.1	Object ID . . . . .	53
3.4.2	Date Time . . . . .	54
3.4.3	Quantity . . . . .	55
3.4.4	Category . . . . .	57
3.4.5	Text . . . . .	58
3.4.6	Sequential Pattern Based Features . . . . .	59
3.5	Mutual Reduction (MR) . . . . .	61
3.5.1	Redundancy Identification . . . . .	63
3.5.2	Mutual Gain Ratio Test . . . . .	64
3.6	Algorithm of Feature Mining . . . . .	65
3.7	Evaluation . . . . .	65
3.7.1	Accuracy Test . . . . .	66

3.7.2	Accumulation Test . . . . .	67
3.7.3	Effectiveness of Feature Reduction . . . . .	68
3.7.4	Effectiveness Comparison . . . . .	69
3.8	Conclusions . . . . .	70
<b>chapter 4</b>	<b>Converging Pattern Mining . . . . .</b>	<b>71</b>
4.1	Introduction . . . . .	71
4.2	Converging Region . . . . .	74
4.3	Basic Definitions . . . . .	76
4.4	Candidates Generation . . . . .	78
4.4.1	Pattern Border . . . . .	78
4.4.2	Subtraction of Pattern Borders . . . . .	79
4.5	Pattern Verification . . . . .	81
4.5.1	T*-tree Index . . . . .	81
4.5.2	Splitting Strategies . . . . .	82
4.6	Algorithm of Converging Pattern Mining . . . . .	85
4.7	Learning Applications for Fraud Detection . . . . .	87
4.7.1	Pattern Selection . . . . .	87
4.7.2	Scoring for Classification . . . . .	88
4.8	Experiment and Evaluation . . . . .	89
4.8.1	Accuracy . . . . .	89
4.8.2	Efficiency . . . . .	91
4.8.3	Effectiveness of The Pruning Strategies . . . . .	92
4.9	Conclusion . . . . .	93
<b>chapter 5</b>	<b>Efficient Globally Optimal Rule Selection . . . . .</b>	<b>94</b>
5.1	Introduction . . . . .	94
5.2	Problem Definition . . . . .	96
5.3	Maximal Coverage Gain Mining . . . . .	100
5.3.1	Pruning Strategies . . . . .	100
5.3.2	Hinge Set Discovery . . . . .	108
5.3.3	Gain Bounding . . . . .	115

5.3.4	Max Coverage Gain Mining: MCGminer . . . . .	117
5.4	Experimental evaluation . . . . .	120
5.4.1	Baseline Settings . . . . .	120
5.4.2	Datsets . . . . .	121
5.4.3	Accuracy Evaluation . . . . .	123
5.4.4	Stability of Detection Rate Against Imbalance Rate . .	123
5.4.5	Effectiveness of Pruning Strategies . . . . .	124
5.4.6	Scalability on Number of Rules . . . . .	125
5.5	Conclusions . . . . .	126
<b>chapter 6</b>	<b>Case Study: Application of Contrast Mining . . .</b>	<b>129</b>
6.1	Case 1: Online Banking Fraud Detection . . . . .	129
6.1.1	Background . . . . .	129
6.1.2	The Framework of Fraud Detection with Contrast Mining	131
6.1.3	Implementation of Alert System: i-Alertor . . . . .	133
6.1.4	Performance Evaluation of i-Alertor . . . . .	135
6.2	Case 2: Key Indicator Analysis in Education . . . . .	138
6.2.1	Background . . . . .	138
6.2.2	Key Indicators Analysis:i-Educator . . . . .	139
6.2.3	Performance Evaluation . . . . .	143
6.3	Conclusion . . . . .	146
<b>chapter 7</b>	<b>Conclusions and Future Work . . . . .</b>	<b>148</b>
7.1	Conclusions . . . . .	148
7.2	Future Work . . . . .	150
7.2.1	Sequential Contrast Pattern Mining in Heterogeneous Data . . . . .	150
7.2.2	Hierarchical Feature Mining in High Dimensional Het- erogeneous Data . . . . .	154
7.2.3	Dynamic Adaptive Anomaly Detection . . . . .	154
<b>chapter A</b>	<b>Appendix: List of Publications . . . . .</b>	<b>156</b>

chapter B	Appendix: List of Symbols . . . . .	158
Bibliography	. . . . .	160



# List of Figures

1.1	The work flow of anomaly detection with contrast mining . . .	4
1.2	Coverage of rules $r_1$ and $r_2$ , where black and white dots represent fraud and genuine transactions, respectively. . . . .	12
1.3	Coverage of $r_1, r_2, r_3$ and $r_4$ , where solid and hollow dots represent fraud and genuine transactions, respectively. . . . .	12
1.4	ROC curve for M_raw and M_derived . . . . .	16
1.5	The profile of the research work of this thesis . . . . .	22
2.1	The framework of literature review . . . . .	24
2.2	Support plane . . . . .	36
3.1	Performance comparison under different feature numbers . . .	46
3.2	Profiling of transaction amount . . . . .	53
3.3	Navigation sequence of a trojan . . . . .	59
3.4	Navigation sequence of a genuine transaction . . . . .	60
3.5	Correlation map . . . . .	63
3.6	Distribution of raw and synthetic attributes . . . . .	66
3.7	Accuracy test among popular classification methods . . . . .	67
3.8	Detection Performance related to the Accumulation of Synthetic Features . . . . .	68
3.9	Effectiveness of Reduction . . . . .	69
3.10	Effectiveness of Reduction . . . . .	70
4.1	Converging region . . . . .	75

4.2	An Example of Converging patterns . . . . .	76
4.3	Framework of building the anomaly detector powered by CPs, where oval boxes display the supporting techniques for each stage . . . . .	78
4.4	An example of T*-tree, where $R_i$ stands for MBB of an internal node . . . . .	83
4.5	ROC on Data set 1 . . . . .	89
4.6	ROC on Data set 2 . . . . .	90
4.7	Efficiency against contrast . . . . .	91
4.8	Effectiveness on strategies . . . . .	92
5.1	Transaction merging, each dot represents a transaction, and the value in every dot is the corresponding gain factor. . . . .	101
5.2	Islands after removing $r'$ . . . . .	102
5.3	Enumeration process . . . . .	107
5.4	Coverage of $R$ on $T'$ . . . . .	109
5.5	Find hinge set when $column_1 = r_2$ . . . . .	113
5.6	Hinge scale test . . . . .	124
5.7	Hinge scale test . . . . .	125
5.8	Splitting balance test . . . . .	126
5.9	The computational cost comparison . . . . .	127
6.1	Framework of Fraud Detection . . . . .	132
6.2	Our fraud detection platform: i-Alertor . . . . .	134
6.3	Fraud Distribution in ExpertSystem and i-Alertor . . . . .	135
6.4	Detection performance comparison between ExpertSystem and i-Alertor . . . . .	137
6.5	Framework of i-Educator . . . . .	140
6.6	The platform of i-Educator . . . . .	142
6.7	Performance evaluation of i-Educator . . . . .	144
6.8	Performance evaluation of i-Educator . . . . .	145
6.9	Some key indicators . . . . .	146

## *LIST OF FIGURES*

---

7.1	A sample of heterogeneous sequence . . . . .	152
7.2	Hierarchical process for feature extraction . . . . .	155

# List of Tables

1.1	Patterns for Fraud Detection . . . . .	5
1.2	An example of task for fraud detection in online banking . . .	7
1.3	Transactions of Online Banking . . . . .	11
1.4	Transactions of Online Banking . . . . .	15
3.1	Correlation matrix $M_c$ . . . . .	46
3.2	Transactions of Online Banking . . . . .	48
3.3	Transactions of Online Banking . . . . .	56
3.4	An example of combinations among multiple “weak” attributes 62	
4.1	Patterns for Fraud Detection . . . . .	71
5.1	An example of task for fraud detection in online banking . . .	94
5.2	Diagonal Matrix of $R$ on $T'$ . . . . .	110
5.3	Matrix with the hinge set as the head . . . . .	111
5.4	The Improvement on Accuracy Evaluation . . . . .	122

# Abstract

Class imbalance data, in which the classes are not equally represented and the minority classes include a much smaller number of examples than other classes, is pervasive and ubiquitous, particularly in applications such as fraud/intrusion detection, medical diagnosis/monitoring, and risk management. The conventional classifiers tend to be overwhelmed by the large classes while ignoring the smaller classes. Typically, many of the existing solutions to the class imbalance problem are proposed at the data level, and a few at the algorithmic level. However, the prior methods have more or less limitations in anomaly detection according to our extensive experiments. Therefore, the thesis targets contrast mining to solve the problem of anomaly detection in imbalanced data from three aspects: feature construction, an effective algorithm for mining contrast patterns, and selection of optimal rule combinations through analysing rule interactions.

Feature construction is one of the most important steps in contrast pattern mining, and any other data mining processes as well. The majority of feature construction methods, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Fourier Transformation, and Independent Component Analysis, usually generate new features by transforming the existing raw features into a new data space. Therefore, previous solutions have many limitations with respect to the objective of training highly accurate classifiers in class imbalance data sets. Incomprehensible features may be generated, based on the assumption that all the samples are independent, the feature set is unstable and sensitive to trivial change of the sample set,

it is difficult to integrate significant domain knowledge, and the classifiers built on the transformed feature set suffer from high False Positive Rate in the class imbalance data set.

In order to train high performance models in the imbalance scenario, we propose a novel method, Personalised Domain Driven Feature Mining (*PDDFM*), to generate important features by integrating domain knowledge effectively with a full consideration of the correlations among samples. A framework specially designed for *PDDFM* is introduced. A novel feature selection method, called Mutual Reduction, is proposed to minimise the noise from redundant features and maximize the contribution of “trivial” features whose gain ratio are low but contribute positively when cooperate with the others. The experimental evaluation reveals our feature mining approach outperforms state-of-the-art methods in anomaly detection.

Contrast pattern mining has been studied intensively for its strong discriminative capability. However, state-of-the-art methods rarely consider the class imbalance problem, which has been proven to be a significant challenge in mining large scale data. The thesis introduces a novel pattern, i.e. converging pattern, which refers to the item sets whose supports contrast sharply from the minority class to the majority class. A novel algorithm, ConvergMiner, is also proposed to mine converging patterns efficiently. A light-weighted index T\*-tree is built to speed up the search process, and output patterns instantly. A series of branch bound pruning strategies are further presented to greatly reduce the computational cost. Substantial experiments on large scale real-life online banking transactions for fraud detection show that the ConvergMiner greatly outperforms the existing cost-sensitive classification methods in terms of accuracy. In particular, it efficiently and effectively detects the frauds in large-scale imbalanced transaction sets. More importantly, the efficiency improves with the increase in data imbalance. After many converging patterns are generated, we propose an effective novel method to select the optimal pattern set.

Rule-based anomaly and fraud detection systems often suffer from sub-

stantial false alerts in the context of a very large number of enterprise transactions with class imbalance characteristics. A crucial and challenging problem is to effectively select a globally optimal rule set which can capture very rare anomalies dispersed in large-scale background transactions. The existing rule selection methods which suffer significantly from complex rule interactions and overlapping in large imbalanced data, often lead to very high false positive rates. We analyse the interactions and relationships between rules and their coverage in transactions, and propose a novel metric, *Max Coverage Gain (MCG)*. *MCG* selects the optimal rule set by evaluating the contribution of each rule in terms of overall performance to cut out those locally significant, but globally redundant rules, without any negative impact on the recall. An effective algorithm, *MCGminer*, is then designed with a series of built-in mechanisms and pruning strategies to handle complex rule interactions and reduce computational complexity in identifying the globally optimal rule set. Substantial experiments on 13 UCI data sets and a real time online banking transactional database demonstrate that *MCGminer* achieves significant improvement in accuracy, scalability, stability and efficiency with respect to large imbalanced data compared to several state-of-the-art rule selection techniques.

Following that, the above proposed contrast analysis techniques have been applied in two industrial projects. The first project was “Fraud Detection in Online Banking” for a major bank in Australia. We developed a risk management platform called *i-Alertor*, which is mainly powered by the techniques introduced in this thesis. According to the evaluation report, *i-Alertor* outperforms the existing rule based system by 10%. The second project was the “Key Indicator Discovery in Student Learning” for a key University in Australia. Another platform called *i-Educator* is also developed to support this application.